Christopher Morgan

CS478 : Brother Christophe

Sep 20, 2013

A. k-Nearest Neighbors.

**1. Design a simple distance metric for this space. Briefly justify your choice.**

D =

Such that:

|  |  |  |
| --- | --- | --- |
| Abbreviation | Attribute Name | Attribute Representation |
| 0 | Outlook | 0 = Rainy  1 = Overcast  2 = Sunny |
| T | Temperature | 0 = Cool  1 = Mild  2 = Hot |
| H | Humidity | 0 = Normal  1 = Hot |
| W | Wind | 0 = Weak  1 = Strong |

Justification:

The thought here is that even though the data is nominal it can by interpreted as ordinal, if not continuous. For example, outlook has an arguable scale to it; conditions are improving as one walks from rainy to overcast to sunny. Overcast is less rainy and more sunny, where sunny is farther away from rainy than overcast, extra. Therefore, the distance between these items holds, as well as their ordering. It is also reasonable that the ratios between these elements hold. With humidity and wind the argument is simply the existence of one or the other. This means it is plausible that a numeric representation can work here for humidity and wind as well.

With the assumption that each attribute can be represented numerically, a Euclidian distance function should give a distance measure between each instance.

**2. Perform 7-fold cross-validation with k=3 for this dataset. Show your work (there should be 7 iterations with 7 corresponding predictions/accuracies on the held-out folds, and a final result).**

|  |  |  |
| --- | --- | --- |
| Run (Instances Used for Testing) | Number Correct (of 2) | Accuracy of Run |
| S1 (1,2) | 0 | 0 |
| S2 (3,4) | 1 | .5 |
| S3 (5,6) | 1 | .5 |
| S4 (7,8) | 2 | 1 |
| S5 (9,10) | 2 | 1 |
| S6 (11,12) | 0 | 0 |
| S7 (13,14) | 1 | .5 |
| Total | **7/14** | **.5** |

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| RUN | Test Instances | 3 Closest Neighbors | Prediction | Number of Correct Predictions in RUN |
| 1 | 1 | 3,8,13 | Y | 0 |
|  | 2 | 3,8,11 | Y |  |
| 2 | 3 | 1,13,2 | N | 1 |
|  | 4 | 14,5,10 | Y |  |
| 3 | 5 | 4,7,10 | Y | 1 |
|  | 6 | 7,14,4 | Y |  |
| 4 | 7 | 6,5,9 | Y | 2 |
|  | 8 | 1,2,3 | N |  |
| 5 | 9 | 7,8,11 | Y | 2 |
|  | 10 | 13,3,4 | Y |  |
| 6 | 11 | 2,7,8 | N | 0 |
|  | 12 | 14,2,3 | N |  |
| 7 | 13 | 10,3,1 | Y | 1 |
|  | 14 | 4,12,6 | Y |  |

I used the following table, which represents the distance from any instance to any other instance.

|  |
| --- |
| 0.00 | 1.00 | 1.00 | 2.24 | 3.00 | 3.16 | 2.65 | 1.00 | 2.24 | 1.73 | 1.73 | 1.73 | 1.41 | 2.45 | |
| 1.00 | 0.00 | 1.41 | 2.45 | 3.16 | 3.00 | 2.45 | 1.41 | 2.45 | 2.00 | 1.41 | 1.41 | 1.73 | 2.24 | |
| 1.00 | 1.41 | 0.00 | 1.41 | 2.45 | 2.65 | 2.45 | 1.41 | 2.45 | 1.41 | 2.00 | 1.41 | 1.00 | 1.73 | |
| 2.24 | 2.45 | 1.41 | 0.00 | 1.41 | 1.73 | 2.00 | 2.00 | 2.45 | 1.41 | 2.45 | 1.41 | 1.73 | 1.00 | |
| 3.00 | 3.16 | 2.45 | 1.41 | 0.00 | 1.00 | 1.41 | 2.45 | 2.00 | 1.41 | 2.45 | 2.00 | 2.24 | 1.73 | |
| 3.16 | 3.00 | 2.65 | 1.73 | 1.00 | 0.00 | 1.00 | 2.65 | 2.24 | 1.73 | 2.24 | 1.73 | 2.45 | 1.41 | |
| 2.65 | 2.45 | 2.45 | 2.00 | 1.41 | 1.00 | 0.00 | 2.00 | 1.41 | 1.41 | 1.41 | 1.41 | 2.24 | 1.73 | |
| 1.00 | 1.41 | 1.41 | 2.00 | 2.45 | 2.65 | 2.00 | 0.00 | 1.41 | 1.41 | 1.41 | 1.41 | 1.73 | 2.24 | |
| 2.24 | 2.45 | 2.45 | 2.45 | 2.00 | 2.24 | 1.41 | 1.41 | 0.00 | 1.41 | 1.41 | 2.00 | 2.24 | 2.65 | |
| 1.73 | 2.00 | 1.41 | 1.41 | 1.41 | 1.73 | 1.41 | 1.41 | 1.41 | 0.00 | 1.41 | 1.41 | 1.00 | 1.73 | |
| 1.73 | 1.41 | 2.00 | 2.45 | 2.45 | 2.24 | 1.41 | 1.41 | 1.41 | 1.41 | 0.00 | 1.41 | 1.73 | 2.24 | |
| 1.73 | 1.41 | 1.41 | 1.41 | 2.00 | 1.73 | 1.41 | 1.41 | 2.00 | 1.41 | 1.41 | 0.00 | 1.73 | 1.00 | |
| 1.41 | 1.73 | 1.00 | 1.73 | 2.24 | 2.45 | 2.24 | 1.73 | 2.24 | 1.00 | 1.73 | 1.73 | 0.00 | 2.00 | |
| 2.45 | 2.24 | 1.73 | 1.00 | 1.73 | 1.41 | 1.73 | 2.24 | 2.65 | 1.73 | 2.24 | 1.00 | 2.00 | 0.00 | |

One important note is that the distance values in my table are rounded to two decimal places. Given how relatively close these data points are it is not unreasonable that extra precision could greatly increase the accuracy of this particular experiment. Running this particular experiment programmatically proved to increase the accuracy by about 10%.

B. Naive Bayes.

**1. Using the full dataset, induce the corresponding NB model. Show your result in the form of probability tables as we did in class.**

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Play Tennis |  |  |  |  |  |  |
|  | **Yes** | **No** |  |  |  |  |
|  | 0.64 | 0.36 |  |  |  |  |
|  |  |  |  |  |  |  |
| Outlook |  |  |  | Humidity |  |  |
|  | **Yes** | **No** |  |  | **Yes** | **No** |
| **Sunny** | 0.22 | 0.60 |  | **High** | 0.33 | 0.80 |
| **Overcast** | 0.44 | 0.00 |  | **Normal** | 0.66 | 0.20 |
| **Rainy** | 0.33 | 0.40 |  |  |  |  |
|  |  |  |  |  |  |  |
| Temperature | |  |  | Wind |  |  |
|  | **Yes** | **No** |  |  | **Yes** | **No** |
| **Hot** | 0.22 | 0.40 |  | **Strong** | 0.33 | 0.60 |
| **Mild** | 0.44 | 0.40 |  | **Weak** | 0.66 | 0.40 |
| **Cool** | 0.33 | 0.20 |  |  |  |  |

**2. What would your model predict for the following two Saturday mornings: <Oct 1, Overcast, Cool, High, Weak>, and <May 26, Sunny, Hot, Normal, Strong>? Show your work.**

|  |  |  |
| --- | --- | --- |
| <Oct 1> | Yes | No |
| Overcast | 0.44 | 0.00 |
| Cool | 0.33 | 0.20 |
| High | 0.33 | 0.80 |
| Weak | 0.66 | 0.40 |
| Yes/No | 0.64 | 0.36 |
| **Product** | **0.02** | **0.00** |
|  |  |  |
| <May 26> | Yes | No |
| Sunny | 0.22 | 0.60 |
| Hot | 0.22 | 0.40 |
| Normal | 0.66 | 0.20 |
| Strong | 0.33 | 0.60 |
| Yes/No | 0.64 | 0.36 |
| **Product** | **0.006** | **0.01** |

For **Oct 1**, my model would predict **Yes**, and for **May 26**, my model would predict **No**.